CSCE 689: Special Topics in Modern Algorithms for Data Science Fall 2023

Lecture 8 — September 8, 2023

Prof. Samson Zhou *Scribe: Wenjing Chen*

1 Low Distortion Embedding

Dimension reduction is an important problem in many fields of machine learning and data analysis. Last time we defined a low distortion embedding as follows:

- Given $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d$, a distance function D, and an accuracy parameter $\epsilon \in [0, 1)$, a low-distortion embedding of $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d$ is a set of points $\mathbf{y}_1, ..., \mathbf{y}_n$ and a distance function D' such that for all $i, j \in [n]$,

$$(1-\epsilon)D(\mathbf{x}_i, \mathbf{x}_j) \leq D'(\mathbf{y}_i, \mathbf{y}_j) \leq (1+\epsilon)D(\mathbf{x}_i, \mathbf{x}_j)$$

In Euclidean space, the distance function D is defined as the l_2 norm of the difference of two points, i.e.,

$$D(\mathbf{x}_i, \mathbf{x}_j) = ||\mathbf{x}_i - \mathbf{x}_j||_2.$$

Some examples of embeddings in Euclidean space are as follows:

1. Suppose $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d$ all lie on the 1^{st}-axis. Let y_i be the first coordinate of $\mathbf{x}_i$. Then $||y_i - y_j||_2 = ||\mathbf{x}_i - \mathbf{x}_j||_2$ for all $i, j \in [n]$. The embedding has no distortion.
2. Suppose $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d$ lie in some k-dimensional subspace V of $\mathbb{R}^d$. Then if we rotate V to coincide with the first k axes of $\mathbb{R}^d$ and set $\mathbf{y}_i$ to be the first k coordinates of $\mathbf{x}_i$, then the embedding has no distortion.

General case: Given $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d$ that lie in general position, does there exist an embedding with no distortion? The answer is NO. However, even in the general case, there exists an embedding with ϵ distortion according to Johnson-Lindenstrauss Lemma.

2 Johnson-Lindenstrauss Lemma

Lemma 1 (Johnson-Lindenstrauss Lemma)**.** *Given* $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \in \mathbb{R}^d$ *and an accuracy parameter* $\epsilon \in [0, 1)$*, there exists a linear map* $\Pi : \mathbb{R}^d \to \mathbb{R}^m$ *with* $m = O(\frac{\log n}{\epsilon^2})$ *so that if* $\boldsymbol{y}_i = \Pi \boldsymbol{x}_i$*, then for all* $i, j \in [n]$*:*

$$(1-\epsilon)||\boldsymbol{x}_i - \boldsymbol{x}_j||_2 \leq ||\boldsymbol{y}_i - \boldsymbol{y}_j||_2 \leq (1+\epsilon)||\boldsymbol{x}_i - \boldsymbol{x}_j||_2$$

.

For $d = 10^{12}$, $n = 10^5$, and $\epsilon = 0.5$, we only require $m \approx 6600$, which demonstrates the effectiveness of this lemma.

We first define the distributional Johnson-Lindenstrauss lemma as follows:

Lemma 2 (Distributional Johnson-Lindenstrauss Lemma)**.** *Given $\Pi \in \mathbb{R}^{m \times d}$ with $m = O(\frac{\log n}{\epsilon^2})$ and each entry drawn from $\frac{1}{\sqrt{m}}\mathcal{N}(0,1)$, let $x \in \mathbb{R}^d$ and suppose $y = \Pi x$. Then with probability at least $1-\delta$,*

$$(1-\epsilon)||\boldsymbol{x}||_2 \leq ||\boldsymbol{y}||_2 \leq (1+\epsilon)||\boldsymbol{x}||_2.$$

To prove Lemma 2, recall that for independent Gaussian random variable $a \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $b \sim \mathcal{N}(\mu_2, \sigma_2^2)$, we have

$$a + b \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Let us denote $\mathbf{y} = (y_1, y_2, ..., y_m)$ and that $\mathbf{x} = (x_1, x_2, ..., x_d)$. By the linear transform $\mathbf{y} = \Pi\mathbf{x}$, it follows that

$$y_i = \frac{1}{\sqrt{m}} \sum_{j=1}^{d} \Pi_{i,j} x_j$$

Then $y_i \sim \mathcal{N}(0, \frac{1}{m}||\mathbf{x}||^2)$. Thus we also have $\mathbb{E}[||\mathbf{y}||^2] = \mathbb{E}[y_1^2 + ... + y_m^2] = ||\mathbf{x}||^2$, which is correct in expectation. In fact, $||\mathbf{y}||^2$ is distributed as Chi-Squared random variable with m degrees of freedom (sum of m squared independent Gaussians). Therefore, we can use the following Chi-Squared Concentration Inequality.

Lemma 3 (Chi-Squared Concentration Inequality)**.** *Let Z be a Chi-squared random variable with m degrees of freedom. Then*

$$Pr[|Z - \mathbb{E}Z| \geq \epsilon\mathbb{E}[Z]] \leq 2e^{-m\epsilon^2/8}.$$

By setting $m = O(\frac{\log(1/\delta)}{\epsilon^2})$, the proof of Lemma 2 follows.

Finally, we prove the Johnson-Lindenstrauss lemma using the distributional Johnson-Lindenstrauss lemma.

Proof of Lemma 1: First, notice that if we define $\mathbf{z}_{i,j} = \mathbf{x}_i - \mathbf{x}_j \in \mathbb{R}^d$ for all $i, j \in [n]$, then what we need to prove is $(1-\epsilon)||\mathbf{z}_{i,j}||_2 \leq ||\mathbf{z}'_{i,j}||_2 \leq (1+\epsilon)||\mathbf{z}_{i,j}||_2$ where $\mathbf{z}'_{i,j} = \Pi\mathbf{z}_{i,j}$. Since there are n vectors of x, there should be $\frac{n(n+1)}{2}$ elements in the set $\{\mathbf{z}_{ij}\}_{i,j\in[n]}$. Therefore, we can invoke the distributional Johnson-Lindenstrauss lemma with failure probability $\delta = O\left(\frac{1}{n^3}\right)$. Taking the union bound over all i, j, we can conclude the proof. □